6 TOTAL REWARD CRITERIA

Eugene A. Feinberg

Abstract: This chapter deals with total reward criteria. We discuss the existence and structure of optimal and nearly optimal policies and the convergence of value iteration algorithms under the so-called General Convergence Condition. This condition assumes that, for any initial state and for any policy, the expected sum of positive parts of rewards is finite. Positive, negative, and discounted dynamic programming problems are special cases when the General Convergence Condition holds.

6.1 INTRODUCTION

This chapter deals with the total reward criterion. This criterion is natural for finite horizon problems and for infinite horizon problems in which the number of steps is not fixed but it is finite along most trajectories. The examples include discounted criteria, which can be interpreted as problems with geometrically distributed horizon, as well as other problems such as sequential statistical procedures, stopping, search, and optimal selection problems.

The analysis of finite horizon models is usually based on the analysis of optimality equations and optimality operators. The value function satisfies the optimality equation and a Markov policy constructed by value iteration procedures is optimal. If optimality cannot be achieved, one can sequentially construct an ε-optimal Markov policy for an N-horizon problem by backward induction. Such a policy is formed by actions for which the reward operators, when applied to the value function at the corresponding step, are (ε/N)-close to the optimal value.

For infinite horizon problems, we can distinguish two distinct approaches. The first approach deals with the analysis of optimality (also called, dynamic programming) equations and operators. The second approach studies probability distributions on the sets of trajectories. In fact, the most interesting results have been achieved by combining these two approaches.

As was observed in the sixties, certain properties of optimality operators, namely contracting and motonicity properties, imply the existence of optimal or nearly optimal policies within natural classes of policies. These properties also imply the convergence of algorithms. In general, dynamic programming operators may not possess these properties. However, if the one-step reward function is uniformly bounded and the nonnegative discount factor is less then one, then the contracting property holds. If rewards are nonnegative (nonpositive) then the value iteration algorithm, applied to the zero terminal value, forms a monotone and therefore convergent sequence. The mentioned three models are called discounted, positive, and negative respectively.

The first comprehensive results were obtained for these three models. Blackwell [8] Denardo [18], Strauch [61] studied discounted models, Blackwell [9], Ornstein [50], and Strauch [61] studied positive models, and Strauch [61] studied negative models. The results differ significantly from one model to another. To illustrate these differences, let us consider the situation when the state space is countable and there are no additional assumptions such as compactness of action sets. For discounted and positive models, for each initial state the supremum of the expected total rewards over the class of all policies is equal to the corresponding supremum over the class of all stationary policies; Blackwell [8, 9]. However, it is not true for negative models; Example 6.5. For discounted models, for any positive constant ε, there exist stationary ε-optimal policies. Such policies may not exist for positive models; Blackwell [9]. However, for positive models there exist stationary εV-optimal policies also called multiplicatively ε-optimal; Ornstein [50]. There are other significant differences between positive, negative, and discounted programming. The differences are so significant that for a long period of time it was even was not clear how to formulate unified results. As the result, all textbooks on dynamic programming and Markov decision processes, that deal with infinite-horizon models with total rewards, consider only positive, negative, and discounted models and deal with them separately; see e.g. Ross [53] and Puterman [52].

In addition to positive, negative, and discounted models, problems with arbitrary reward functions and without discounting have been considered in the literature for a long period of time. It turned out that the most comprehensive results can be proved when the so-called General Convergence Condition holds. This condition means that the positive part of the reward function satisfies the positive programming assumptions. Positive, negative, and discounted models are particular cases of models satisfying the General Convergence Condition.

Blackwell [7] and Krylov [45] proved the existence of stationary optimal policies for MDPs with finite state and action sets. Dubins and Savage [20] and Hordijk [43] described necessary and sufficient conditions for optimality, so-called conserving and equalizing conditions. Derman and Strauch [19] and Strauch [61] proved that, for a given initial state, any policy can be substituted with an equivalent randomized Markov policy. Krylov [45] and Gikhman and Skorohod [40] showed that nonrandomized policies are as good as randomized policies. Dynkin and Yushkevich [21] proved that if someone randomizes between different policies, the objective function cannot be improved. Feinberg [22, 23] proved that, for a given initial state, (nonrandomized) Markov policies are as good as general ones; earlier van Hee [70] showed that the supremum of the expected total rewards over the class of Markov policies is equal to the supremum over the class of all policies.

Seminal papers by Blackwell [8, 9] and Strauch [61] dealt with models with Borel state and action spaces. Some of the following papers on total reward MDPs dealt just with countable state spaces. For some results, their extension from countable to uncountable models is a straightforward exercise. For other results, such extensions are either difficult or impossible. In addition, Blackwell and Strauch considered Borel-measurable policies and discovered that ε-optimal policies may not exist but for any initial measure they exist almost everywhere. In order to establish the existence of everywhere ε-optimal policies, one should expand the set of policies to universally measurable or analytically measurable policies. Such extension was introduced by Blackwell, Freedman, and Orkin [11] and it was done in a systematic and comprehensive way in the book by Bertsekas and Shreve [5]. In particular, this book expanded in a natural way almost all results on positive, negative and discounted programming in a way that ε-optimality was established instead of almost sure ε-optimality. The major exception was Ornstein's theorem [50]. It was proved by Ornstein [50] for a countable state space; see also Hordijk [43]. It's extension to Borel models in the sense of almost sure multiplicative nearly-optimality was formulated by Blackwell [9] as an open question. Frid [38] solved this problem (Schäl and Sudderth [59] found a correctable gap in Frid's proof). The natural conjecture is that under more general measurability assumptions, in the spirit of Bertsekas and Shreve [5], almost sure nearly-optimality can be replaced with nearly optimality everywhere in Ornstein's theorem. Blackwell and Ramachandran [12] constructed a counter-example to this conjecture.

For the General Convergence Condition, significantly deeper results are available for countable state models than for Borel state problems. First, three important particular results were discovered by van der Wal: (i) the supremum of the expected total rewards over all policies is equal to the supremum over stationary policies if the action sets are finite; [66, Theorem 2.22] (this result was generalized by Schäl [56] to Borel state models with compact action sets); (ii) extension of Ornstein's theorem to models in which rewards may be nonpositive and in each state, where the value function is nonpositive, a conserving action exists; [68] (this results was generalized by van Dawen and Schäl [65, 64]); (iii) existence of uniformly nearly-optimal Markov policies; [67]. The survey by van der Wal and Wessels [69] describes these and many preceding results.

Feinberg and Sonin [34] generalized Ornstein's [50] theorem to models satisfying the General Convergence Condition. For the long period of time, it had not been clear even how to formulate such results. The first clue is that in the more general formulation the value function of the class of stationary policies should be considered instead of the value function of the class of all policies. The second clue is that, in the definition of multiplicative ε-optimal policies (or in the definition of εV-optimal policies according to another terminology), the value function V should be replaced with an excessive majorant of a value function of the class of stationary policies. The proofs in Feinberg and Sonin [34] are non-trivial and differ from Ornstein's [36] proofs. Feinberg and Sonin [36] extended these results to non-stationary policies and to more general classes of functions that approximate optimal values.

Feinberg [25, 26] described the structure of uniformly nearly-optimal policies in countable state models satisfying the General Convergence Condition. As mentioned above, stationary policies can be significantly outperformed by non-stationary policies in negative problems; see Example 6.5. It turned out that this example demonstrates the only pathological situation when the value of the class of stationary policies is less than the value of the class of all policies. Consider the set of states at which the value function equals zero and there are no conserving actions (actions at which the optimality operator applied to the value function achieves its maximum). Then there are policies which are uniformly nearly optimal and which are stationary outside of this set; see Theorems 6.20 and 6.21.

Another important feature of the papers by Feinberg and Sonin [36, 25, 26] is that they consider general classes of nonstationary policies and general methods how to deals with nonstationary policies. In particular, the information about the past plays an important role. It is possible to identify two properties of this information: (i) Non-Repeating and (ii) Transitivity Conditions. The Non-Repeating condition implies that randomized policies are as good as non-randomized ones and there exist uniformly nearly optimal policies within any class of policies that satisfies this condition. The Transitivity Condition implies that the model can be transformed into a new model in a way that the class of policies satisfying this condition in the old model becomes the class of stationary policies in the new one.

This paper is a survey of results and methods for models satisfying the General Convergence Condition. We consider countable state MDPs everywhere in this paper, except Section 6.10. We present the theory for countable state MDPs, discuss Borel state MDPs, and discuss open questions, most of which deal with uncountable state spaces. In order to illustrate major concepts and counter-examples, we start our presentation with classical discounted, positive, and negative problems.

6.2 DEFINITIONS OF DISCOUNTED, POSITIVE, NEGATIVE, AND GENERAL CONVERGENT MODELS

We say that an MDP is *discounted* if function r is bounded and there is a constant $\beta \in [0,1[$, called the discount factor, such that

$$v(x,\pi) = \mathbb{E}_x^\pi \sum_{t=0}^{\infty} \beta^t r(x_t, a_t). \tag{6.1}$$

An MDP is called *unbounded discounted* if (6.1) holds and the function r is bounded above, $r(x,a) \le C < \infty$ for all $x \in \mathbb{X}$, $a \in \mathbb{A}(x)$ and for some C.

Discounted and unbounded discounted MDPs can be reduced to an MDP with a discount factor equal to 1. In order to do it, we add an additional state to the state space $\mathbb{X}$. This state has only one action under which it is absorbent and all rewards are equal to zero in this state. This state sometimes is called a grave. The one-step transition probabilities between states in $\mathbb{X}$